Datamining\_HW\_32121391  
  
2.1 아래의 문제를 해결하기 위해 데이터마이닝 기법을 이용하려고 한다. 지도학습과 비지도학습 중 어느 기법을 사용하는 것이 보다 적절한지 구분하시오.  
  
a. 과거 소비자들의 인구통계학적, 금융 데이터를 바탕으로 새로운 소비자들의 대출여부를 결정  
-> 지도학습(기존 데이터 통해 소비자 대출여부 결정(정답여부 존재))  
  
b. 온라인 서점에서 기존 구매 고객들의 구매 패턴을 파악하여 책을 추천  
-> 비지도학습(비슷한 데이터를 군집화, (정답여부 존재 X))  
  
c. 기존에 알고 있는 네트워크 데이터 패킷 상태 정보를 이용하여 새로운 네트워크 데이터 패킷의 상태 분류(예: 바이러스, 해커 공격)  
-> 지도학습(기존 데이터를 통해 새로운 데이터 패킷 분류(정답여부 존재))  
  
d. 비슷한 성향을 가지고 있는 고객군 파악  
-> 비지도학습(비슷한 성향 데이터 군집화 (정답여부 존재X))  
  
e. 파산한 회사와 그렇지 않은 회사들의 재정상태 데이터를 기반으로 특정 회사의 파산 여부를 예측  
-> 지도학습(파산여부를 파악해야 함(정답여부 존재))  
  
f. 수리이력 정보를 바탕으로 항공기 수리 시간을 추정  
-> 지도학습(예상 수리 시간 추정(정답여부 존재))  
  
g. 우편번호 스캔 결과 정보를 바탕으로 우편물 자동분류  
-> 비지도학습(우편번호는 해당 단지까지 만을 알려주지 정확한 호수까지 알려주지 않는다.  
배달부는 해당 지역의 우편물을 여러 개 받는 것이기에 군집화로 여겨진다. 고로(정답여부 존재X))  
  
h. 본인의 구매물품과 다른 고객들의 과거 구매 이력 정보를 바탕으로 할인 쿠폰의 발행 여부 판단  
-> 지도학습(쿠폰을 발행할지 안 할지 결정(정답여부 존재))

#### 2.11

setwd("C:/Users/ycg00/Desktop/2019-1/데이터마이닝")  
Toyota <- read.csv("ToyotaCorolla.csv", header = T)  
str(Toyota)

## 'data.frame': 1436 obs. of 39 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Model : Factor w/ 319 levels "TOYOTA Corolla ",..: 276 276 276 276 275 275 269 269 257 246 ...  
## $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  
## $ Age\_08\_04 : int 23 23 24 26 30 32 27 30 27 23 ...  
## $ Mfg\_Month : int 10 10 9 7 3 1 6 3 6 10 ...  
## $ Mfg\_Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  
## $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  
## $ Fuel\_Type : Factor w/ 3 levels "CNG","Diesel",..: 2 2 2 2 2 2 2 2 3 2 ...  
## $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  
## $ Met\_Color : int 1 1 1 0 0 0 1 1 0 0 ...  
## $ Color : Factor w/ 10 levels "Beige","Black",..: 3 7 3 2 2 9 5 5 6 3 ...  
## $ Automatic : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CC : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  
## $ Doors : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ Cylinders : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Gears : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Quarterly\_Tax : int 210 210 210 210 210 210 210 210 100 185 ...  
## $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...  
## $ Mfr\_Guarantee : int 0 0 1 1 1 0 0 1 0 0 ...  
## $ BOVAG\_Guarantee : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Guarantee\_Period : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ ABS : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_1 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_2 : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Airco : int 0 1 0 0 1 1 1 1 1 1 ...  
## $ Automatic\_airco : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Boardcomputer : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ CD\_Player : int 0 1 0 0 0 0 0 1 0 0 ...  
## $ Central\_Lock : int 1 1 0 0 1 1 1 1 1 0 ...  
## $ Powered\_Windows : int 1 0 0 0 1 1 1 1 1 0 ...  
## $ Power\_Steering : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Radio : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Mistlamps : int 0 0 0 0 1 1 0 0 0 0 ...  
## $ Sport\_Model : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ Backseat\_Divider : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Metallic\_Rim : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Radio\_cassette : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Parking\_Assistant: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Tow\_Bar : int 0 0 0 0 0 0 0 0 0 0 ...

##### (a) 시각화 기능을 이용해 상관성이 큰 변수쌍을 찾으시오.

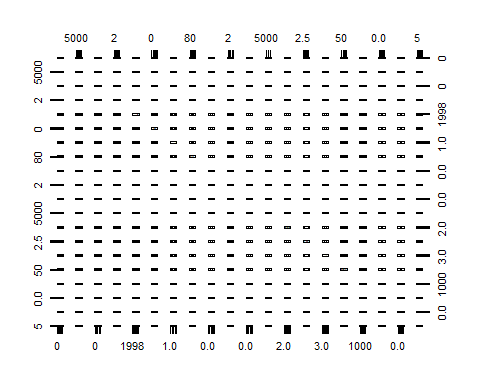
우선 dim()를 통해 데이터셋의 크기를 봤다.

dim(Toyota)

## [1] 1436 39

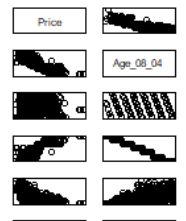
###### 그냥 모든 데이터셋을 plot()에 넣을 경우 디스플레이 크기 제한으로볼 수 없다. 데이터셋을 보니 21번째 열 넘어서는 부품 데이터만 나열 2열부터 21번째 열까지만 상관도를 본다.

plot(Toyota[,2:21])



plot 함수를 통해 (Zoom을 통해 모니터 상에서 확인했다.)확인한 결과  
Price ~ Age\_08\_04 + Mfg\_Year +KM +Weight  
Age\_08\_04 ~ Mfg\_Year + KM

변수 정도가 강한 상관관계를 갖고 있는 것을 확인했다.



<- 확대본 sample

##### **(b )****i - Fuel\_Type 변수와 Metallic 변수는 범주형이다. 어떤 방식으로 이 변수들을 가변수로 변환할지 설명하시오. R을 이용하여 범주형 변수를 가변수로 변환하시오.**

###### 우리가 바꿔야 하는 범주형 변수 Metalic, Fuel\_Type

###### str(Toyota)로 확인했을 때 Metallic은 이미 전환이 끝났다.

###### Fuel\_Type 를 바꿔본다.

Ch\_Fuel\_Type <- model.matrix(~0 + Toyota$Fuel\_Type, data = Toyota)  
Ch\_Fuel\_Type <- as.data.frame(Ch\_Fuel\_Type)  
str(Ch\_Fuel\_Type)

## 'data.frame': 1436 obs. of 3 variables:  
## $ Toyota$Fuel\_TypeCNG : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Toyota$Fuel\_TypeDiesel: num 1 1 1 1 1 1 1 1 0 1 ...  
## $ Toyota$Fuel\_TypePetrol: num 0 0 0 0 0 0 0 0 1 0 ...

t(t(names(Ch\_Fuel\_Type)))

## [,1]   
## [1,] "Toyota$Fuel\_TypeCNG"   
## [2,] "Toyota$Fuel\_TypeDiesel"  
## [3,] "Toyota$Fuel\_TypePetrol"

head(Ch\_Fuel\_Type)

## Toyota$Fuel\_TypeCNG Toyota$Fuel\_TypeDiesel Toyota$Fuel\_TypePetrol  
## 1 0 1 0  
## 2 0 1 0  
## 3 0 1 0  
## 4 0 1 0  
## 5 0 1 0  
## 6 0 1 0

###### Model 변수와 Fuel\_Type 변수를 지우고 새로 만든 변수들을 집어넣자

Toyota\_NEW <-cbind(Toyota[,-c(2,8)],Ch\_Fuel\_Type)  
str(Toyota\_NEW)

## 'data.frame': 1436 obs. of 40 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  
## $ Age\_08\_04 : int 23 23 24 26 30 32 27 30 27 23 ...  
## $ Mfg\_Month : int 10 10 9 7 3 1 6 3 6 10 ...  
## $ Mfg\_Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  
## $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  
## $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  
## $ Met\_Color : int 1 1 1 0 0 0 1 1 0 0 ...  
## $ Color : Factor w/ 10 levels "Beige","Black",..: 3 7 3 2 2 9 5 5 6 3 ...  
## $ Automatic : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CC : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  
## $ Doors : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ Cylinders : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Gears : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Quarterly\_Tax : int 210 210 210 210 210 210 210 210 100 185 ...  
## $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...  
## $ Mfr\_Guarantee : int 0 0 1 1 1 0 0 1 0 0 ...  
## $ BOVAG\_Guarantee : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Guarantee\_Period : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ ABS : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_1 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_2 : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Airco : int 0 1 0 0 1 1 1 1 1 1 ...  
## $ Automatic\_airco : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Boardcomputer : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ CD\_Player : int 0 1 0 0 0 0 0 1 0 0 ...  
## $ Central\_Lock : int 1 1 0 0 1 1 1 1 1 0 ...  
## $ Powered\_Windows : int 1 0 0 0 1 1 1 1 1 0 ...  
## $ Power\_Steering : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Radio : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Mistlamps : int 0 0 0 0 1 1 0 0 0 0 ...  
## $ Sport\_Model : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ Backseat\_Divider : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Metallic\_Rim : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Radio\_cassette : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Parking\_Assistant : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Tow\_Bar : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Toyota$Fuel\_TypeCNG : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Toyota$Fuel\_TypeDiesel: num 1 1 1 1 1 1 1 1 0 1 ...  
## $ Toyota$Fuel\_TypePetrol: num 0 0 0 0 0 0 0 0 1 0 ...

###### 마지막 변수 3가지를 보면 추가한 변수임을 확인할 수 있다.

###### + 덧붙임. 다르게 만든 거

nnrow<-dim(Toyota)[1]  
nnrow

## [1] 1436

levels(Toyota$Fuel\_Type)

## [1] "CNG" "Diesel" "Petrol"

###### Fuel\_Type은 3가지의 변수가 있었다. Diesel, Petrol, CNG 이들을 각각의 벡터로 나누어, 그곳에 맞는 값들만 1이 형성되도록 만들 거다. 우선 더미 변수를 만들어 공간을 확보한다.

dummy\_D<-rep(0,nnrow)  
dummy\_P<-rep(0,nnrow)  
dummy\_C<-rep(0,nnrow)

###### 빈 공간인 더미 변수에 각 변수 위치에 존재하는 숫자를 넣는다.

D\_index <- which(Toyota$Fuel\_Type=="Diesel")  
P\_index <- which(Toyota$Fuel\_Type=="Petrol")  
C\_index <- which(Toyota$Fuel\_Type=="CNG")

###### 더미 파일은 지금 빈 상태. index는 몇 번째 해당 단어가 있는지 알려주는 것이므로 해당 index 마다 1을 넣는다고 하면

dummy\_D[D\_index] <-1  
dummy\_P[P\_index] <-1  
dummy\_C[C\_index] <-1

###### 가변수로 만든 CNG, Diesel, Petrol 함수 3가지를 합치면…

Fuel <- data.frame(dummy\_C,dummy\_D, dummy\_P)  
names(Fuel) <- c("Fuel\_CNG","Fuel\_Diesel","Fuel\_Petrol")  
str(Fuel)

## 'data.frame': 1436 obs. of 3 variables:  
## $ Fuel\_CNG : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Fuel\_Diesel: num 1 1 1 1 1 1 1 1 0 1 ...  
## $ Fuel\_Petrol: num 0 0 0 0 0 0 0 0 1 0 ...

###### Fuel 이라는 데이터 프레임 안에 CNG, Diesel, Petrol 컬럼이 있다. 이 데이터를 원 데이터인 Toyota와 묶는다. 묶는 김에 id, Model, Fuel\_Type 변수를 없앤다.

###### id = 1, Model=2, Fuel\_Type = 8 번째니까

Toyota\_New <- cbind(Toyota[,-c(1,2,8)], Fuel)  
str(Toyota\_New)

## 'data.frame': 1436 obs. of 39 variables:  
## $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  
## $ Age\_08\_04 : int 23 23 24 26 30 32 27 30 27 23 ...  
## $ Mfg\_Month : int 10 10 9 7 3 1 6 3 6 10 ...  
## $ Mfg\_Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  
## $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  
## $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  
## $ Met\_Color : int 1 1 1 0 0 0 1 1 0 0 ...  
## $ Color : Factor w/ 10 levels "Beige","Black",..: 3 7 3 2 2 9 5 5 6 3 ...  
## $ Automatic : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ CC : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  
## $ Doors : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ Cylinders : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Gears : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Quarterly\_Tax : int 210 210 210 210 210 210 210 210 100 185 ...  
## $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...  
## $ Mfr\_Guarantee : int 0 0 1 1 1 0 0 1 0 0 ...  
## $ BOVAG\_Guarantee : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Guarantee\_Period : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ ABS : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_1 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Airbag\_2 : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Airco : int 0 1 0 0 1 1 1 1 1 1 ...  
## $ Automatic\_airco : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Boardcomputer : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ CD\_Player : int 0 1 0 0 0 0 0 1 0 0 ...  
## $ Central\_Lock : int 1 1 0 0 1 1 1 1 1 0 ...  
## $ Powered\_Windows : int 1 0 0 0 1 1 1 1 1 0 ...  
## $ Power\_Steering : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Radio : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Mistlamps : int 0 0 0 0 1 1 0 0 0 0 ...  
## $ Sport\_Model : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ Backseat\_Divider : int 1 1 1 1 1 1 1 1 0 1 ...  
## $ Metallic\_Rim : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Radio\_cassette : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Parking\_Assistant: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Tow\_Bar : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Fuel\_CNG : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Fuel\_Diesel : num 1 1 1 1 1 1 1 1 0 1 ...  
## $ Fuel\_Petrol : num 0 0 0 0 0 0 0 0 1 0 ...

###### **(b) –ii** **// R을 이용하여 이 데이터를 학습데이터(50%), 검증데이터(30%), 평가데이터(20%)로 분할하고 각각의 역할에 대해 설명하시오. (분할시 모든 변수를 사용하고 초기값은 디폴트값을 사용하시오)**

###### 우선 set.seed를 통해 sampling 과정을 고정한다.

set.seed(1234)

###### 데이터를 나눠야 한다.

###### 학습데이터(50%): 모델을 구축 시 사용되는 데이터를 의미한다. trn<- sample(rownames(Toyota\_New), dim(Toyota\_New)[1]\*0.5)) trn.data <-Toyota\_New[trn,]

검증데이터(30%): 학습데이터로부터 구축된 여러 모델의 성능을 비교할 때 사용한다.  
val<- sample(setdiff(rownames(Toyota\_New), trn),dim(Toyota\_New)[1]\*0.3))  
val.data <-Toyota\_New[val,]

평가데이터(20%): 구축된 모델이 미래에 수집될 새로운 데이터에 대해 얼마만큼 예측/분류 성능을 보일지 평가할 때 쓰인다.  
test <-setdiff(rownames(Toyota\_New),union(trn,val))  
test.data<- Toyota\_New[test,]

이건 따로 찾아본 거  
  
학습데이터(50%)  
trn\_tyt<-sample(1:nnrow, round(0.5\*nnrow))  
Toyota\_Trn<- Toyota\_New[trn\_tyt,]

검증데이터(30%)  
 val\_tyt<- sample(setdiff(1:nnrow, trn\_tyt), round(0.3\*nnrow))  
Toyota\_Val<- Toyota\_New[val\_tyt,]  
  
평가데이터(20%)  
test\_tyt<- setdiff(1:nnrow, union(trn\_tyt,val\_tyt))  
Toyota\_Test<-Toyota\_New[test\_tyt,]

**HW2\_DataMining**

**수업: 데이터마이닝**

**강의자: 이성임 교수**

**학번: 32121391**

**학과: 커뮤니케이션학부**

**이름: 박동민**